Overlap Graph and Clumps

Mireille Régnier

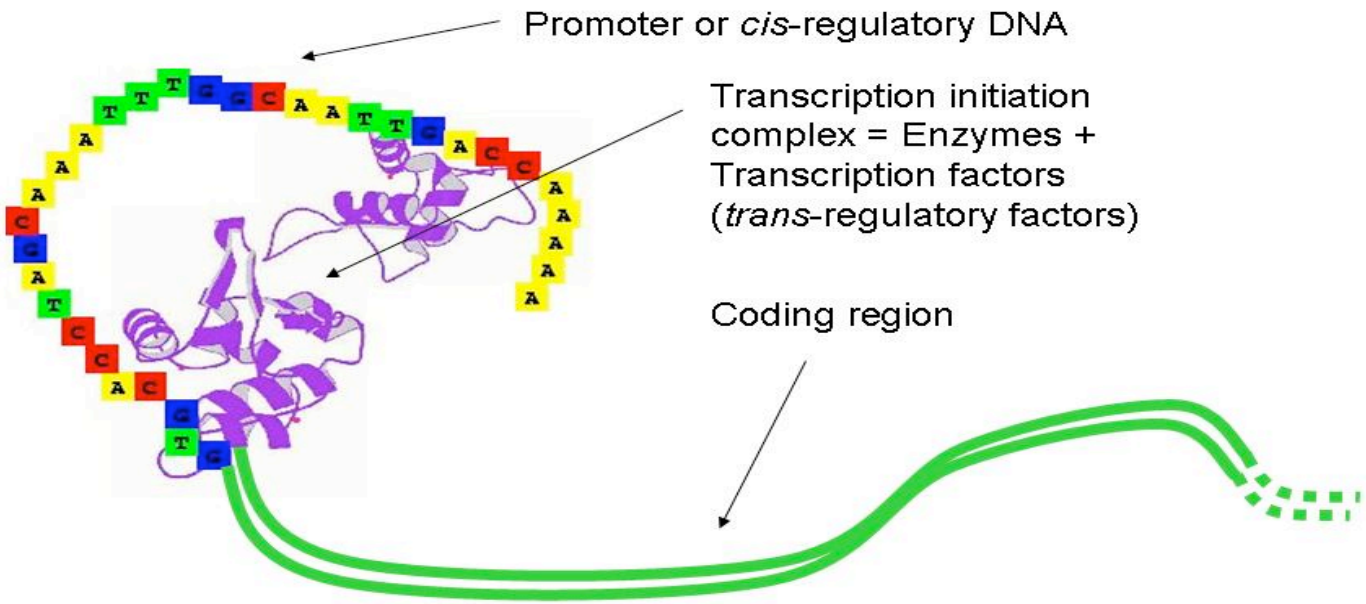
LIX and INRIA
Mireille.Regnier@inria.fr
web page : algo.inria.fr/regnier

October, 9-th – 2008

Outline

Cis-regulation

The control of gene expression results from DNA/Protein interaction in the vicinity of the transcription start



Cis-regulation changes

Arabidopsis thaliana

Arabidopsis lyrata

Arabidopsis halleri

Example : the caudal motif in early developmental enhancers from *Drosophila*

Papatsenko et al., 2002

```
GCTTTTTTATGGTCGGC
TCGCTTTTATGGCCCAA
CAGTTTTTATGTCTTTA
CCGTTTTGATGGCGGTG
AAATTTTTAGGGAACCA
GCCCGTTTATGGTTCCC
GACACTTTATGTGACAA
TCGGATTTATGACACAA
ATGTCTTTATGATTATT
GCAACTTTTGGGCCATA
CCCTTTTGTTGGCCAAA
```

(a) Aligned Motifs

A|	2	3	2	2	1	0	0	0	9	0	0	2	1	3	3	4	7
C|	3	7	3	2	3	0	0	0	0	0	0	0	6	4	5	2	2
G|	4	0	5	1	1	0	0	2	0	2	11	7	1	1	2	1	1
T|	2	1	1	6	6	11	11	9	2	9	0	2	3	3	1	4	1

(b) Countings

Example : the caudal motif in early developmental enhancers from *Drosophila*

Papatsenko et al., 2002

```
GCTTTTTTATGGTCGGC
TCGCTTTTATGGCCCAA
CAGTTTTTATGTCTTTA
CCGTTTTGATGGCGGTG
AAATTTTTAGGGAACCA
GCCCGTTTATGGTTCCC
GACACTTTATGTGACAA
TCGGATTTATGACACAA
ATGTCTTTATGATTATT
GCAACTTTTGGGCCATA
CCCTTTTGTTGGCCAAA
```

(a) Aligned Motifs

A|	2	3	2	2	1	0	0	0	9	0	0	2	1	3	3	4	7
C|	3	7	3	2	3	0	0	0	0	0	0	0	6	4	5	2	2
G|	4	0	5	1	1	0	0	2	0	2	11	7	1	1	2	1	1
T|	2	1	1	6	6	11	11	9	2	9	0	2	3	3	1	4	1

(b) Countings

A|	-0.22	0.06	-0.22	-0.22	-0.62	-1.32	-1.32	-1.32	0.98	-1.32	-1.32	-0.22	-0.62	0.06	0.06	0.28	0.75
C|	0.06	0.75	0.06	-0.22	0.06	-1.32	-1.32	-1.32	-1.32	-1.32	-1.32	-1.32	0.62	0.28	0.47	-0.22	-0
G|	0.28	-1.32	0.47	-0.62	-0.62	-1.32	-1.32	-0.22	-1.32	-0.22	1.16	0.75	-0.62	-0.62	-0.22	-0.62	-0
T|	-0.22	-0.62	-0.62	0.62	0.62	1.16	1.16	0.98	-0.22	0.98	-1.32	-0.22	0.06	0.0 6	-0.62	0.28	-0

(c) Position Specific Scoring matrix

Probability Weight Matrices

Probability function

- Threshhold s : A word (site) is *similar* iff $score(w) > s$.
- Pvalue :

$$Prob_n(\exists H; score(H) > s) \quad .$$

Probability Weight Matrices

Probability function

- Threshhold s : A word (site) is *similar* iff $score(w) > s$.
- Pvalue :

$$Prob_n(\exists H; score(H) > s) \quad .$$

Algorithms and data structures

- candidates-motifs extraction

Probability Weight Matrices

Probability function

- Threshhold s : A word (site) is *similar* iff $score(w) > s$.
- Pvalue :

$$Prob_n(\exists H; score(H) > s) \quad .$$

Algorithms and data structures

- candidates-motifs extraction

Model accuracy

- Improve PWM with structural information

Principles

Biological function

- Overrepresented words
- underrepresented words

Statistical softwares

- candidates-motifs extraction
- statistical significance

Probability Computation

"Classic" methods vs Graphs

- induction ; [GuOd81]
- languages [ReSz98] ;automata [NiFlSa00].

Probability Computation

"Classic" methods vs Graphs

- induction ; [GuOd81]
- languages [ReSz98] ;automata [NiFlSa00].

Space/time complexity

- Exact (all n) $\rightarrow$ AhoPro (NIIGenetika, Inria)
- $O(n \times |\Sigma|)$; n : text size ; Σ : data structure.

Probability Computation

"Classic" methods vs Graphs

- induction ; [GuOd81]
- languages [ReSz98] ;automata [NiFISa00].

Space/time complexity

- Exact (all n) $\longrightarrow$ AhoPro (NIIGenetika, Inria)
- $O(n \times |\Sigma|)$; n : text size ; Σ : data structure.

Drawback

- n dependency ;
- numerical precision ;

Probability Computation

"Classic" methods vs Graphs

- induction ; [GuOd81]
- languages [ReSz98] ;automata [NiFlSa00].

Probability Computation

"Classic" methods vs Graphs

- induction ; [GuOd81]
- languages [ReSz98] ;automata [NiFlSa00].

Space/time complexity

- Approximation $\rightarrow$ RSA-tools, Spatt, AhoSoft (NIIGenetika, Inria)
- $O(1 \times |\Sigma|)$

Probability Computation

"Classic" methods vs Graphs

- induction ; [GuOd81]
- languages [ReSz98] ;automata [NiFlSa00].

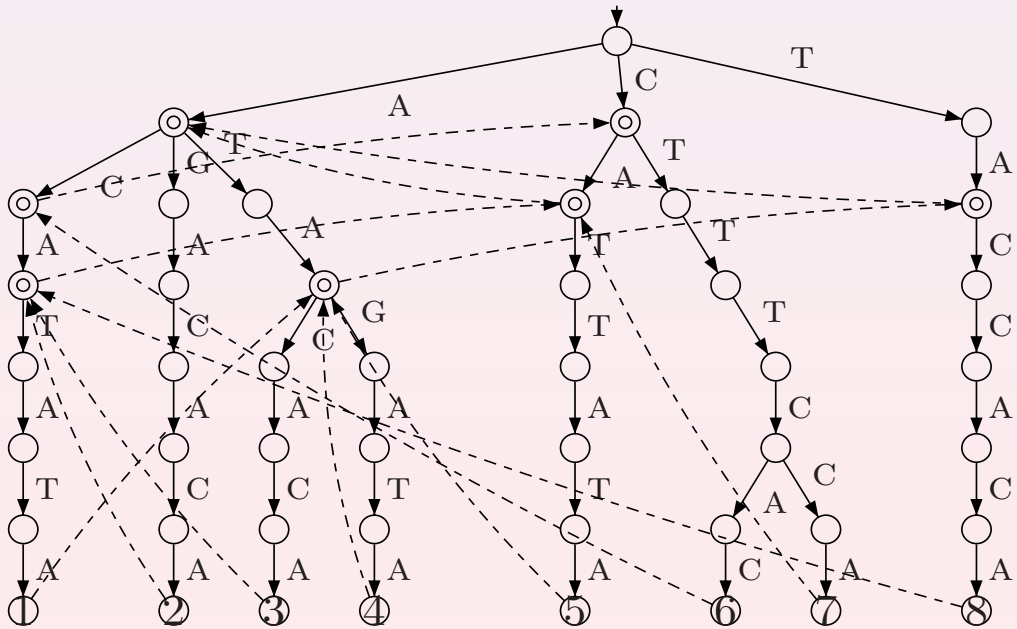
Space/time complexity

- Approximation $\rightarrow$ RSA-tools, Spatt, AhoSoft (NIIGenetika, Inria)
- $O(1 \times |\Sigma|)$

Drawback

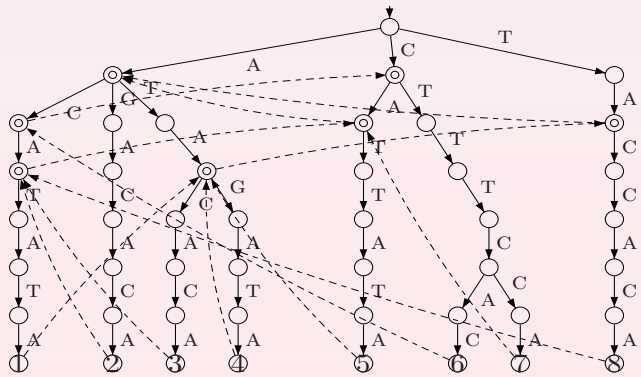
- size of the data structure ;
- tightness ;

AhoCorasick searching automaton



AhoCorasick automaton : searching and computing

- $n : w_n$ = largest prefix found =ATA ;
- $n+1$: character x found :
  - $x = G, wx = ATAG \in Graph, w_{n+1} = ATAG$
  - $x = A, C, T, wx \notin Graph$
    - $x = C; w = A \cdot TA, w_{n+1} = TAC \in Graph$
    - $x = T; w = AT \cdot A, w_{n+1} = AT \in Graph$
    - $x = A; AA, TAA \notin G, w_{n+1} = root$



AhoPo :pobability computation

Step n : $(p_n(w))_{w \in Graph}$.
$p_n(w) = Prob(\text{largest prefix ending at } n \text{ is} w)$.

Induction

$$
\begin{aligned}
p_{n+1}(ATAG) &= p_n(ATA) \cdot p(G) \\
p_{n+1}(AT) &= p_n(ATA) \cdot p(T) \\
&+ p_n(AGA) \cdot p(T) \\
&+ p_n(CA) \cdot p(T) + p_n(TA) \cdot p(T)
\end{aligned}
$$

AhoCorasick automaton : searching and computing

Left relation

$$\mathrm{H}_1 \mathcal{R}_\mathrm{L} \mathrm{H}_2 \quad \Leftrightarrow \quad \mathit{Father}_{\mathit{LOG}}(\mathrm{H}_1) = \mathrm{Father}_{\mathrm{LOG}}(\mathrm{H}_2)$$

$$\{\mathit{ATACACA}, \mathit{ATAGATA}\} \qquad \tilde{\mathit{ATA}}$$

ATA :Largest prefix of ATACACA that is a suffix in $\mathcal{H}$

AhoCorasick automaton : searching and computing

Left relation

$$\mathrm{H}_1 \mathcal{R}_\mathrm{L} \mathrm{H}_2 \quad \Leftrightarrow \quad \mathit{Father}_{\mathit{LOG}}(\mathrm{H}_1) = \mathrm{Father}_{\mathrm{LOG}}(\mathrm{H}_2)$$

$\{\mathit{ATACACA}, \mathit{ATAGATA}\}$ $\quad$ $\mathit{A\tilde{T}A}$

ATA :Largest prefix of ATACACA that is a suffix in $\mathcal{H}$

Right relation

$$\mathrm{H}_1 \mathcal{R}_\mathrm{R} \mathrm{H}_2 \quad \Leftrightarrow \quad \mathit{Mother}_{\mathit{ROG}}(\mathrm{H}_1) = \mathrm{Mother}_{\mathrm{ROG}}(\mathrm{H}_2)$$

$\{\mathit{ATACACA}, \mathit{ATACACA}\}$ $\quad$ $\mathit{A\bar{C}A}$

$\cup\{\mathit{AGACACA}, \}$

ACA :Largest suffix of ATACACA that is a prefix in $\mathcal{H}$

Computation on Graph :induction

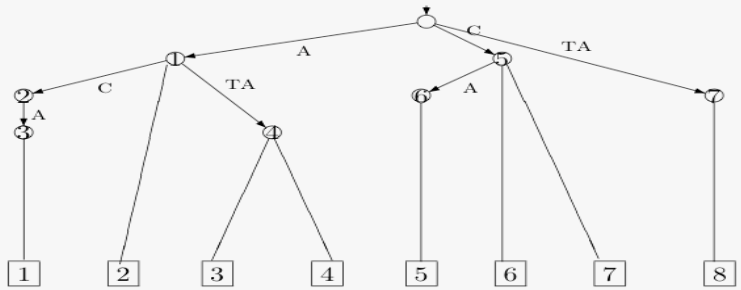


Figure 2: Left-Overlap Graph

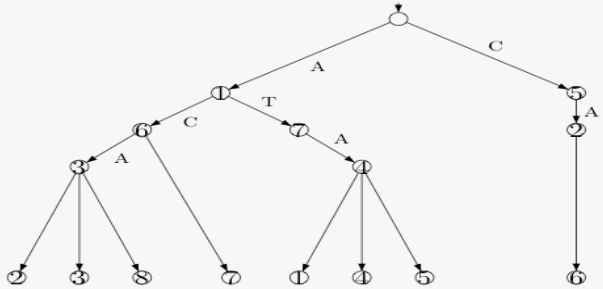


Figure 3: Right-Overlap Graph

AhoCorasick automaton : searching and computing

First occurrence at position $n = 18$

$$\begin{array}{rccc} GGGGGGGG & |ATACACA & | & \\ \text{no } \mathrm{H} \in \mathcal{H} & | & \cdots & |n \end{array}$$

AhoCorasick automaton : searching and computing

First occurrence at position $n = 18$

$$\begin{array}{rll} GGGGGGGG & |ATACACA & | \\ \text{no } \mathrm{H} \in \mathcal{H} & | \quad \cdots & |n \end{array}$$

AND NOT

$$\begin{array}{ll} GGGGCATT| & ATACACA| \\ GGGGACAT| & ATACACA| \end{array}$$

$$\begin{array}{ll} GGACATAT| & ATACACA| \\ GGAGACAC| & ATACACA| \end{array}$$

$$\cdots$$

All marked nodes in AhoGraph

Ovelap graph :pobability computation

Compute $(p_n(\mathrm{H}))_{\mathrm{H}\in\mathcal{H}}$ using LOG, ROG.

LOG

dependency to the past

ROG

information to transfer (memory)

Ovelap graph :pobability computation

Compute $(p_n(\mathrm{H}))_{\mathrm{H}\in\mathcal{H}}$ using LOG, ROG.

LOG

dependency to the past

ROG

information to transfer (memory)

Graph traversals...

Clump counts

- First occurrence : "small" n.
- k occurrences : large n.
- $\Rightarrow$ approximation
- $\Rightarrow$ generating functions
- $\Rightarrow$ clumps

Clump counts

With $\mathrm{H}_1 = \mathrm{AACGGAA}$ and $\mathrm{H}_2 = \mathrm{GAATCA}$,

$$AACGGAACGGAACGGAATCACGGAA$$

k-decomposition counted with coef. $(-1)^k$ [BoCIReVa05].

Clump counts

With $H_1 = AACGGAA$ and $H_2 = GAATCA$,

$$AACGGAACGGAACGGAATCACGGAA$$

k-decomposition counted with coef. $(-1)^k$ [BoCIReVa05].

Contribution $(-1)^7 = -1$

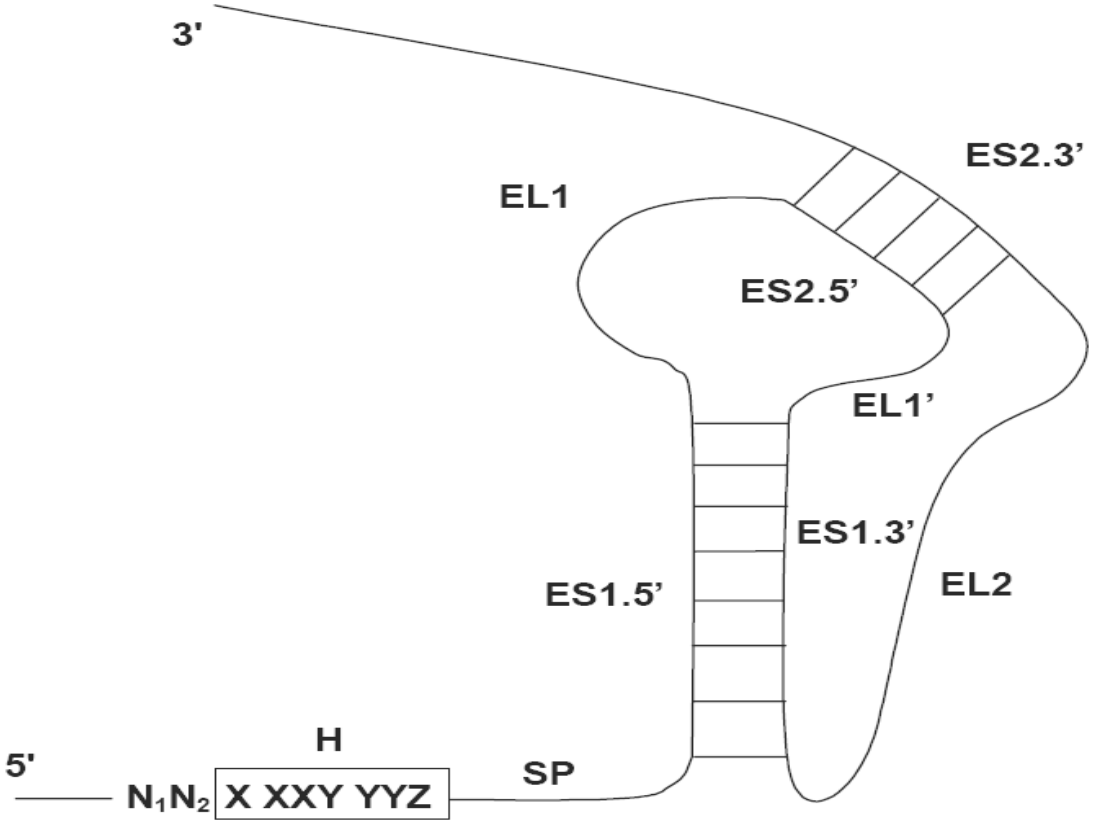
With $AACAACAACAA = AA(CAA)^3$

$$AACAACAACAA{\cdot}CAA{\cdot}CAA{\cdot}CAA{\cdot}CAA{\cdot}CAA{\cdot}CAA{\cdot}CAA{\cdot}$$
$$CAA{\cdot}CAA{\cdot}CAA{\cdot}CAA{\cdot}CAA{\cdot}ACAACAACAA{\cdot}$$

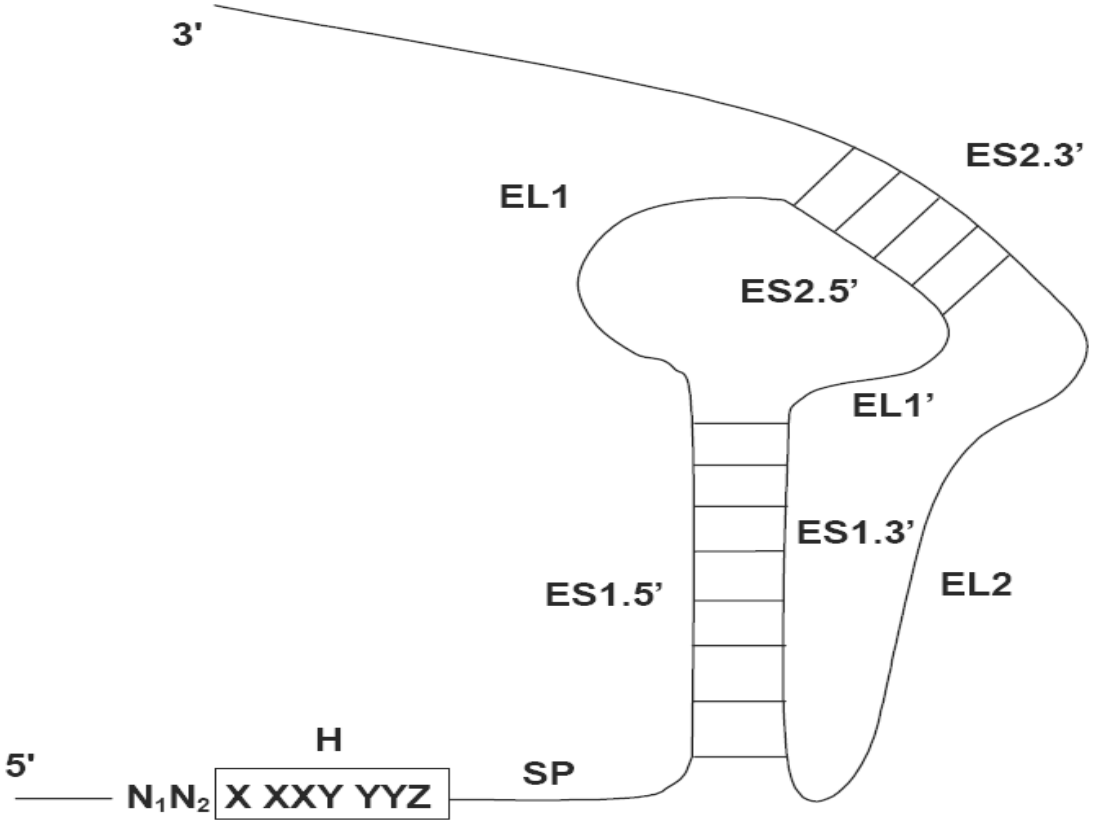
No contribution : even = odd

$$AACAACAACAA{\cdot}CAA{\cdot}CAA{\cdot}CAA{\cdot}CAA{\cdot}CAA{\cdot}CAA{\cdot}CAA{\cdot}$$
$$CAA{\cdot}CAA{\cdot}CAA{\cdot}CAA{\cdot}ACAACAACAA{\cdot}$$

Open problems : Frameshift and riboswitches



Open problems : Frameshift and riboswitches



Boxes :

$$(w_1, w_2, \tilde{w}_1, \tilde{w}_2)$$

with : P. Nicodeme.

Open problems : Frameshift and riboswitches